Example of Central Limit Theorem using Exponential Distribution

José Rodrigo Cervantes Polanco

Overview

The **Central Limit Theorem** (CLT) is one of the most important theorems in statistics. For our purposes, the CLT states that the distribution of averages of I.I.D. random variables becomes that of a normal distribution as the sample size increases. In this project, we will use the exponential distribution as an example to understand the Central Limit Theorem.

Central Limit Theorem (CLT)

Let $\{X_1, \cdots, X_n\}$ be a sequence of I.I.D random variables of size n drawn from distributions of expected values given by μ and finite variance given by σ^2. Suppose we are interested in the sample average

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

of these random variables. The CLT states that as n gets larger, the distribution of $\bar{X}_n$ is close to the normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

Basic settings

We fixed a few basic settings to be used during this analysis.

```
setwd("~/Documents/ProgramsGitHub/StatInference")  # working directory
library(ggplot2)      # to be able to graph
set.seed(0)           # to be able to reproduce
```

Simulations:

In this section we are going to do 1000 simulations of averages of 40 exponentials with $\lambda = 0.2$.

```
numsim = 1000                                                  # Number of simulations
sample = NULL                                                  # Initialize the simulations
for (i in 1 : numsim) sample = c(sample, mean(rexp(40,0.2))) # Create the simulations
sample<-as.data.frame(sample)                                  # Create a dataframe (ggplot)
head(sample)                                                   # First simulations
```

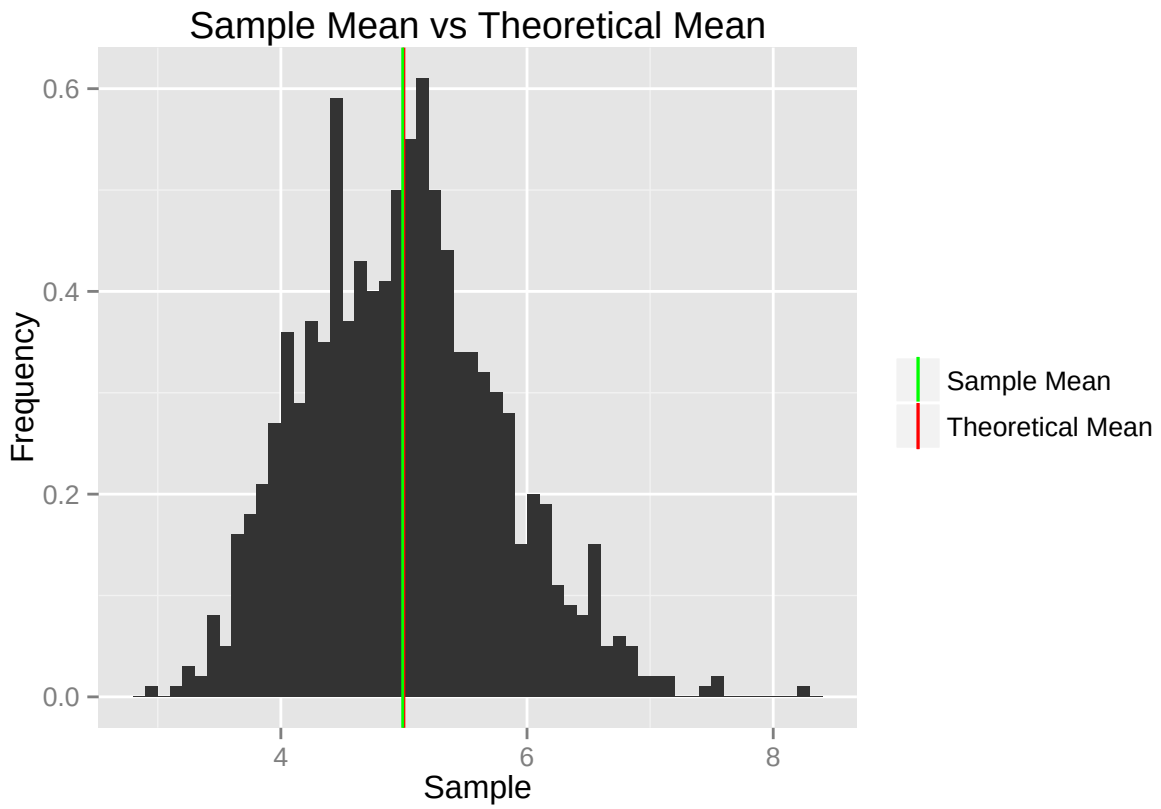
```
##   sample
## 1  4.776
## 2  5.875
## 3  4.260
## 4  4.864
## 5  5.019
## 6  4.579
```

Sample Mean versus Theoretical Mean

Now, we are going to compare the sample mean versus the theoretical mean. We know that the mean of a random variable X with exponential distribution, is given by $\mu(X) = 1/\lambda$. In our simulations $\lambda = 0.2$, therefore $\mu(X) = 5$. A consequence of the CLT is that the mean of the random variable given by the average of 40 exponentials is approximated by the mean of the exponential distribution, in this project 5.

Fisrt, we graph both the sample mean and the theoretical mean:

```
g <- ggplot(sample,aes(sample))     # Enter the database to be graph
g <- g + geom_histogram(binwidth=.1,aes( y=..density..))  # Type of graph / histogram
g <- g + geom_vline(aes(xintercept=5, color="Theoretical Mean"),
                    linetype="solid", size=0.5, show_guide = TRUE)
g <- g + geom_vline(aes(xintercept=mean(sample), color="Sample Mean"),
                    linetype="solid", size=0.5,show_guide = TRUE)
g <- g + scale_colour_manual("", values = c("Sample Mean" = "green",
                                            "Theoretical Mean" ="red" ))
g <- g + labs(title = "Sample Mean vs Theoretical Mean",x="Sample",y="Frequency")
g
```



Second, we calculate the sample mean:

```
mean(sample$sample)
```

```
## [1] 4.99
```

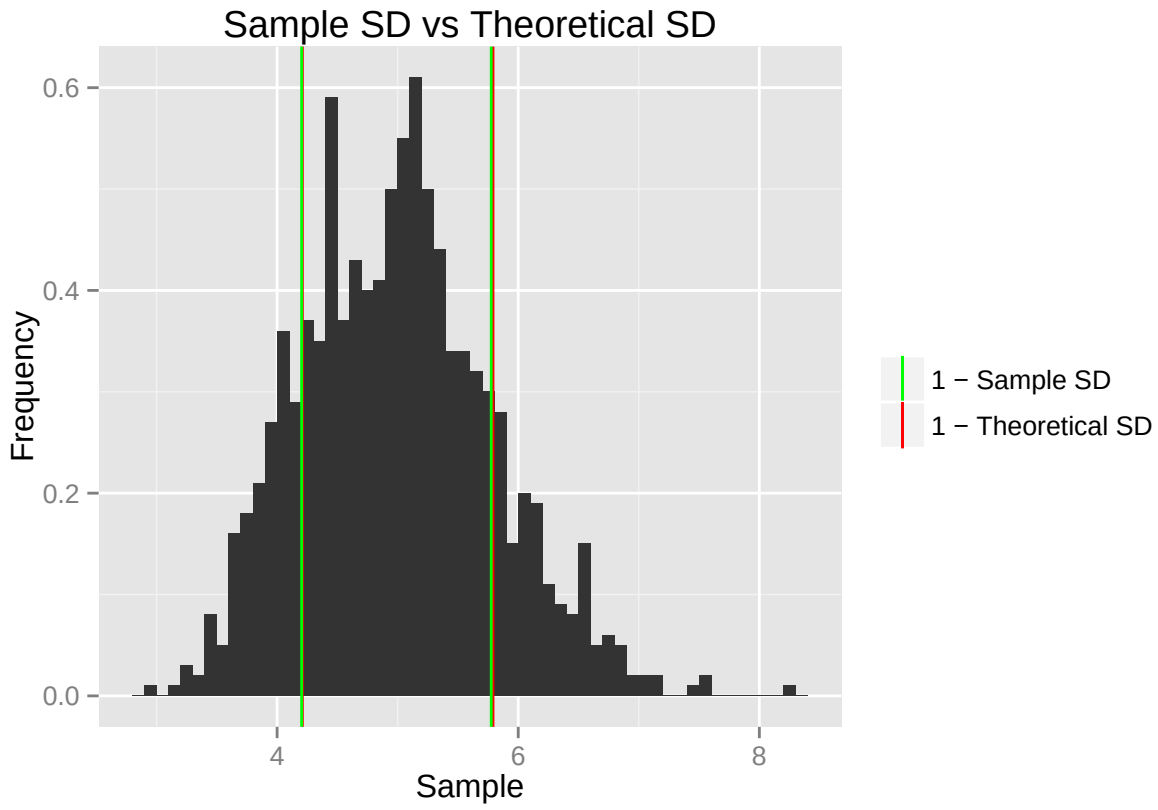
From this calculation and the previous graph, we have checked that 5 is a good approximation for the sample mean.

Sample Variance versus Theoretical Variance:

In this part, we are going to compare the sample variance versus the theoretical variance. We know that the variance of a random variable X with exponential distribution, is given by $\text{var}(X) = 1/\lambda^2$. In our simulations $\lambda = 0.2$, therefore $\text{var}(X) = 25$. A consequence of the CLT is that the variance of the random variable given by the average of 40 exponentials is approximated by the variance of the exponential distribution divide by 40, in this project $25/40 = 0.625$.

Fisrt, we graph bands with a width of 1 sample standard deviation and 1 theoretical standard deviation (remember that the standard deviation is the square root of the variance):

```
g <- ggplot(sample,aes(sample))     # Enter the database to be graph
g <- g + geom_histogram(binwidth=.1,aes( y=..density..))  # Type of graph / histogram
g <- g + geom_vline(aes(xintercept=5+(25/40)^(1/2), color="1 - Theoretical SD"),
                    linetype="solid", size=0.5, show_guide = TRUE)
g <- g + geom_vline(aes(xintercept=5-(25/40)^(1/2), color="1 - Theoretical SD"),
                    linetype="solid", size=0.5, show_guide = TRUE)
g <- g + geom_vline(aes(xintercept=mean(sample)+sd(sample), color="1 - Sample SD"),
                    linetype="solid", size=0.5,show_guide = TRUE)
g <- g + geom_vline(aes(xintercept=mean(sample)-sd(sample), color="1 - Sample SD"),
                    linetype="solid", size=0.5,show_guide = TRUE)
g <- g + scale_colour_manual("", values = c("1 - Sample SD" = "green",
                                            "1 - Theoretical SD" ="red" ))
g <- g + labs(title = "Sample SD vs Theoretical SD",x="Sample",y="Frequency")
g
```



Second, we calculate the sample mean:

```r
var(sample$sample)
```

```
## [1] 0.6182
```

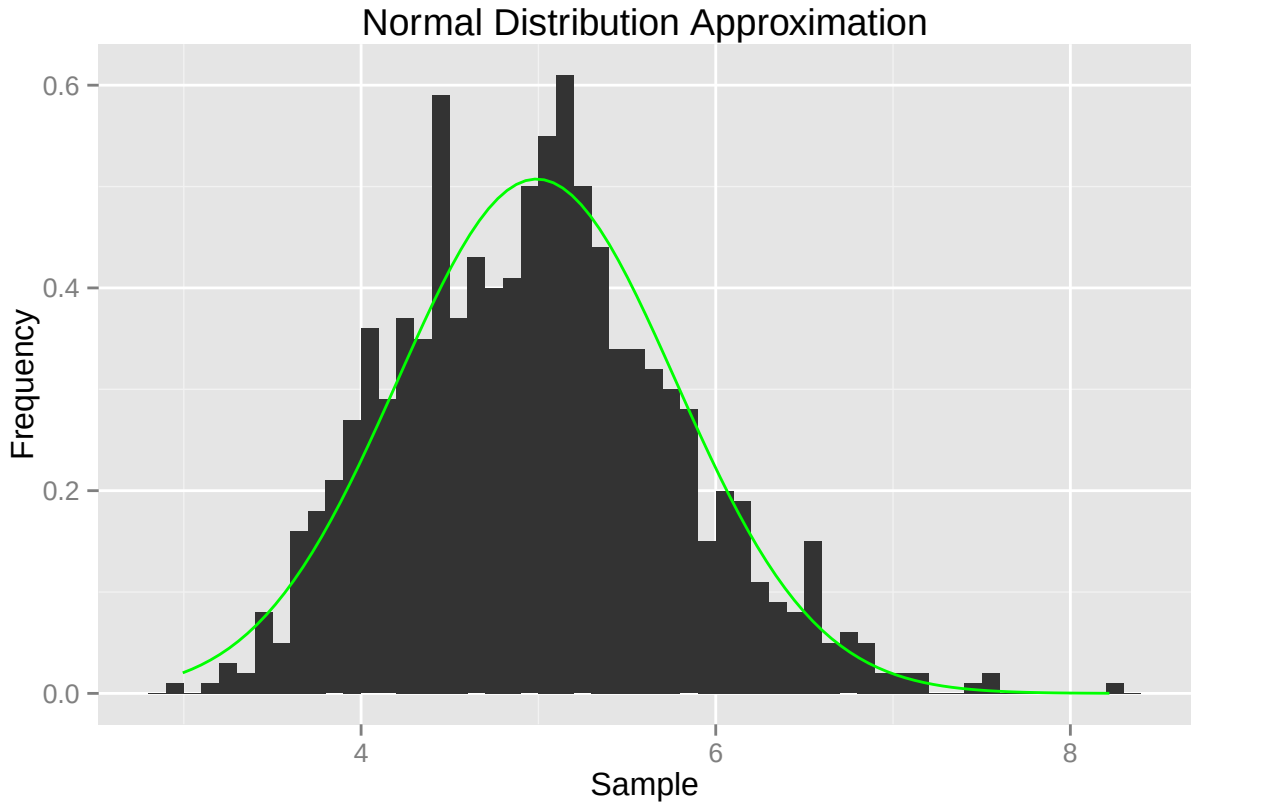
From this calculation and the previous graph, we have checked that $25/40 = 0.625$ is a good approximation for the sample variance.

Distribution:

The way to think about the CLT is that $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$, where $\bar{X}_n$ is the average of I.I.D. random variables, n the number of random variables, μ is the mean and σ the variance of one of those random variables. In our

Finally, we graph a normal distribution with mean equals 4.99 and variance equals 0.6182 over the histogram of the frequencies of the sample that we have been using in this project:

```r
g <- ggplot(sample,aes(sample))      # Enter the database to be graph
g <- g + geom_histogram(binwidth=.1,aes( y=..density..))  # Type of graph / histogram
g <- g + stat_function(fun = dnorm,args = list(mean=mean(sample$sample),
                                               sd = sd(sample$sample)), color = "green")
g <- g + labs(title="Normal Distribution Approximation",x="Sample",y="Frequency")
g
```



The last graph shows that the distribution of our sample of 1000 averages of 40 exponentials with $\lambda = 0.2$ can be approximated by a normal distribution with mean 4.99 (approx 5) and variance 0.6182 (approx 0.625).